NOT TO BE FURTHER COPIED OR REPRODUCED

(* BATCHES *)

# 6

## Skipping Analyses

(u-, t-, and s-values)
[may be delayed]


| | | Exh. | Page |
|---|---|---|---|
| Introduction 6-1 | | a | |
| **6A.** Setting values aside 6-2 | | b 1 | 6-5/7 |
| | | 2 | 6-9 |
| **6B.** An introductory example 6-4 | | c 3 | 6-12 |
| **6C.** Batches of 3 and 4 6-10 | | d 4 | 6-13/15 |
| **6D.** Continuing the processes; Cushny and Peebles 6-11 | | 5 | 6-16 |
| | | 6 | 6-18 |
| **6E.** Standard patterns of skipping; u-, t-, and s- 6-19 | | e | |
| **6F.** Further examples 6-21 | | f 7 | 6-22 |
| - hydroelectric power again 6-21-6-29 | | 8 | 6-23 |
| - atomic weights 6-29 to 6-31 | | 9 | 6-25 |
| | | 10 | 6-26 |
| **6G.** Generations in data analysis 6-31 | | 11 | 6-27/28 |
| | | 12 | 6-30 |
| ********** | | g | |
| **6H.** How far have we come? 6-33 | | h | |
| | | k | |
| ********** | | l 13 | 6-36 |
| | | 14 | 6-37 |
| **6K.** How do the skipping midmeans and trimeans respond to changing one value in a batch? (background) 6-35 | | m 15 | 6-39 |
| | | 16 | 6-40 |
| **6L.** Skipmeans (optional technique) 6-35 | | p 17 | 6-43 |
| **6M.** s-means 6-38 | | 18 | 6-44 |
| | | 19 | 6-45 |
| ********** | | | |
| **6P.** Problems 6-41 | | | |

# 6

## Skipping Analyses

## (u-, t-, and s- values)

It might seem that we went far enough when we chose simple middle means as our standard central values. For some purposes -- notably the calculation of residuals within single batches -- this is usually true. There the need is to use "fitted values" that are:

- at about the right level.
- are not violently disturbed by the value of any one observation.

Trimeans and midmeans meet these conditions with ease.

We do -- and will -- use central values for other purposes, however. We want to use them to summarize batches -- and to help define residuals that are to be compared with residuals from other fitted values. For such purposes, we are likely to want to use even better central values than trimean or midmean -- if we can do this without too much difficulty.

For such purposes, further improvements are well worthwhile.

Why do we need further improvement? Because "stray" observations still influence our central values, not quantitatively -- their exact value does not matter to a trimean or a midmean -- but directionally -- the side of the other observations on which they lie affects the ranks of all observations, and thus the detailed values of hinges, etc.

The simplest way to meet this difficulty is to do another analysis with identifiably stray values unranked. We shall say that such values have been

set aside

This does not mean that they have been discarded. Not at all. They have merely been given special treatment.

Part of this special treatment -- whenever possible -- should be looking back at the data as hard and carefully as possible. How are these observations identified? What special circumstances, if any, distinguish them from others? What remarks seem relevant?

Sometimes the major value of a set-aside procedure lies in the answers these questions provoke.

## 6A. Setting values aside

To set aside some values but not others requires a dividing line. We saw in the last chapter that the dividing lines of exploratory data analysis need not be sharp -- that, in fact, it was better that they were somewhat hazy.

The detective art is a tentative one. The finding of clues and the evaluation of their properties may seem -- may sometimes even be -- a matter of knowledge definitely gained. But this is not typical. The detective, like the bloodhound on a confused trail, must much more often cast about for new paths, for new alternative -- and often very tentative -- ways to think about his problem.

As we hone the edges of our tools for cutting into data, we must expect their use to become more an art -- and hence,in particular, more tentative. The form of report

illustrated in the last chapter -- where we reported some points as detached, but others as only outside, -- was not an accident.

Wherever feasible, we expect to behave in a similar way -- making two different settings-aside, looking at two (or more) different analyses -- in most cases where setting aside is to be considered.

The fact that setting aside a value can affect one analysis but not the other is important. Setting aside a value that were better kept is not likely to be catastrophic, if there is one analysis in which it is retained. As a result we can afford to be rather free and unworried about setting values aside.

To be sure, doing this places a greater burden on the human interpreter. This is, however, just as it should be. The proper purpose of carefully organized processes -- especially those suitable for automation -- is to provide results that can be usefully subjected to human insight and judgment.

If we are to have various levels of setting aside, we are likely to need more than one cutting value. The last chapter's side values and three-halves values provide us with two choices, both of which will prove to be useful.

Depending on what we think of the problem, we may want to look at either:

- two analyses -- one with no set aside the other with some.

- two analyses -- with varying amounts of set aside.

## 6B. An introductory example

Exhibit 1 takes up a batch of 25 values and illustrates the calculation in various ways.
We find one outside observation high and one detached observation low. We have now to proceed to an analysis of the batch after we have set aside:

- only the detached observations.

- both the outside and the detached observations.

The compressed notations toward the foot of the panel C are modified in several ways:

- that for the initial analysis is marked "U" -- for "unskipped" -- opposite the count.

- that for the skipping analysis using three-halves points as cutting values and setting aside detached points only is marked "T1" opposite the count.

- that for the skipping analysis using side values as cutting values and setting aside both outside and detached points is marked "S1" opposite the count.

- the additional section introduced in the last chapter is included, giving the difference of hinges ("dh"), the side and three-halves values ("S/T"), and lists or counts of observations respectively below and above the corresponding side values and three-halves values ("out").

exhibit 1 of chapter 6

Side points and outside values for a familiar batch of 25 values

**A)** The DATA

5, -52, -27, -83, 8, -14, -122, -110, 112, 58, -119, 33, 18, -52, -19, 12, -82, 14, 25, -182, -40, 64, -56, 5, 13.

**B)** The ARITHMETIC

Hinges: -56 and +14

Midspread: 70 = 14 - (-56)

1/2 Midspread 35 = 1/2(70)

Side Values: -126 = -56 - 70

+84 = +14 + 70

Three-Halves Values = -161 = -126 - 35

119 = +84 + 35

Outside Points: +112

Detached Points: -182

(Cont'd)

## 6C. Batches of 3 and 4

Somehow, very small batches are always a special case. In particular, so long as we lack extrasensory perception, there is no reasonable way, when there are 1 or 2 values in a batch, to look only at the batch and set aside any values.

In a batch of 3, the extremes are always the hinges. As a consequence, the side values, which are outside the hinges, are outside the extremes. If we follow the pattern set in the last chapter, we can never find any outside point. The easy modification is to call an extreme "outside" if it is more than 10 times as far from the nearest value as that value is from the farthest value. We shall adopt this procedure for batches of 3 values. (There is nothing sacred about 10. It is easy to remember.)

The procedure of the last chapter will also find no outside values starting with a batch of 4 values. Here the simplest reasonable rule calls an extreme "outside" if it is more than twice as far from the nearest value as that nearest value is from the farthest value. (Again, nothing sacred about "twice", but much convenience, both in memory and use.)

These definitions correspond to simple definitions of side values:

- in batches of 3, each side value is ten times as far from the median as the opposite extreme.

- in batches of 4, each side value is twice as far from the adjacent middle value as is the opposite extreme.

Exhibit 3 illustrates the arithmetic.

6D. Continuing the processes; Cushny and Peebles

Our use of setting aside could consist of just what we did in exhibit 1.

It could also consist of repeating that process just as many times as new detached and outside points are found. To do this, we set aside the points found in one step before repeating the process on the remaining points in the next step.

Exhibit 4 does this in all detail for a famous set of data. The first step of the S analysis causes 4.6 and 0.0 to be set aside. Repeating the process for the eight remaining observations causes 2.4 to be set aside. Repeating the process for the third time -- for the seven remaining observations -- has no effect. Accordingly the S-process stops.

The first step of the T-analysis sets aside only one value, 4.6. The second step sets aside two other values, 0.0 and 2.4. The third step sets aside nothing new, stopping the T-process.

As one becomes familiar with the calculation, it is not necessary to write down all the detail shown in exhibit 4. Exhibit 5 shows a streamlined version of the same calculation.

exhibit 3 of chapter 6

Side values and outside values for batches of 3 or 4

A) BATCH of 3

Data: -7, -3, 122

Side points: $-3 + 10(-3 - (-7)) = 37$

$-3 - 10(122 - (-3)) = -1253$

Outside value: 122

B) Another EXAMPLE

Data: -7, 2, 5

Side points: $2 + 10(2 - (-7)) = 92$

$2 - 10(5-2) = -28$

Outside value: none

C) BATCH of 4

Data: -7, -3, 2, 122

Side points: $2 + 2(2 - (-7)) = 20$

$-3 - 2(122 - (-3)) = -253$

Outside value: 122

D) Another EXAMPLE

Data: -7, -3, 2, 5

Side points: $-3 - 2(5 - (-3)) = -19$

$2 + 2(2 - (-7)) = 20$

Outside value: none

exhibit 4 of chapter 6

Increase in additional hours of sleep, drug 2
as Compared with drug 1 in 10 patients

A) ALL 10 VALUES -- initial step of both analyses

| | | | | |
|---|---|---|---|---|
| 4.6 | 10 | U | dh | 0.8 up 0.4 |
| 2.4 | 3H | 1.0 | S/T | 0.2 / -0.2 |
| 1.8 H | 3H | 1.8 | S/T | 2.6 /3.0 |
| 1.4 | | | out | 0.0 / none |
| 1.3 | | | out | none / 4.6 |
| 1.3 | | | | |
| 1.2 | | | | |
| 1.0 H | | | | |
| 0.8 | | | | |
| 0.0 | | | | |

B) 9 REMAINING VALUES -- analyzed for second step of T-analysis

| | | | | |
|---|---|---|---|---|
| 2.4 | 9 | T1 | dh | 0.4 up 0.2 |
| 1.8 | 3H | 1.0 | T | 0.4 |
| 1.4 H | 4H | 1.4 | T | 2.0 |
| 1.3 | | | det | 0.0 |
| 1.3 | | | det | 2.4 |
| 1.2 | | | | |
| 1.0 H | | | | |
| 0.8 | | | | |
| 0.0 | | | | |

C) 7 REMAINING VALUES -- analyzed for third step of T-analysis

| | | |
|---|---|---|
| | 7 | T2 |
| 1.8 | 3H | 1.0 |
| 1.4 H | 4H | 1.4 |
| 1.3 | | |
| 1.3 | | |
| 1.2 | | |
| 1.0 H | | |
| 0.8 | | |

(hinges the same as T1, so T-process stops)

(Cont'd)

exhibit 4 of chapter 6 (Cont'd)

D) FINAL VERSION of T-ANALYSIS

| | | | |
|---|---|---|---|
| 4.6 | | 7 | T = T2 |
| 2.4 | | H | 1.0 |
| 1.8 | | M | 1.3 |
| 1.4 | H | H | 1.4 |
| 1.3 | | tri | 1.25 |
| 1.3 | M | | |
| 1.2 | | | |
| 1.0 | H | | |
| 0.8 | | | |
| 0.0 | | | |

E) 8 REMAINING VALUES -- analyzed for second step of S-analysis

| | | | | | |
|---|---|---|---|---|---|
| 2.4 | | 8 | S1 | dh | 0.5 |
| 1.8 | (H1.6) | 3.5H | 1.1 | S | 0.6 |
| 1.4 | | 3.5H | 1.6 | S | 2.1 |
| 1.3 | | | | out | no more |
| 1.3 | | | | out | 2.4 |
| 1.2 | | | | | |
| 1.0 | (H1.1) | | | | |
| 0.8 | | | | | |

F) 7 REMAINING VALUES -- analyzed for third step of S-analysis

| | | | | | |
|---|---|---|---|---|---|
| 1.8 | | 7 | S2 | dh | 0.4 |
| 1.4 | H | 3H | 1.0 | S | 0.6 |
| 1.3 | | 4H | 1.4 | S | 1.8 |
| 1.3 | | | | out | no more |
| 1.2 | | | | out | no more |
| 1.0 | H | | | | |
| 0.8 | | | | | |

(no more outside values, so S-process stops)

G) FINAL VERSION of S-ANALYSIS

| | | | |
|---|---|---|---|
| 4.6 | | 7 | S = S2 |
| 2.4 | | H | 1.0 |
| 1.8 | | M | 1.3 |
| 1.4 | H | H | 1.4 |
| 1.3 | | tri | 1.25 |
| 1.3 | M | | |
| 1.2 | | | |
| 1.0 | H | | |
| 0.8 | | | |
| 0.0 | | | |

(Cont'd)

exhibit 4 of chapter 6 (Cont'd)

H) ALTERNATE FORM of SCRAWL -- s-analysis as an example

| | |
|---|---|
| 4 | 6 |
| 3 | |
| 2 | 4 |
| 1 | 023348 |
| 0 | 08 |

| | |
|---|---|
| 10 | U |
| 3H | 1.0 |
| 3H | 1.8 |
| dh | 0.8 |
| S | 0.2 |
| S | 2.6 |
| out | (1) |
| out | (1) |

| | |
|---|---|
| 8 | S1 |
| 3.5H | 1.1 |
| 3.5H | 1.6 |
| dh | 0.5 |
| S | 0.6 |
| S | 2.1 |
| out | (1) |
| out | (2) |

| | |
|---|---|
| 7 | S2 |
| 3H | 1.0 |
| 4H | 1.4 |
| dh | 0.4 |
| S | 0.6 |
| S | 1.8 |
| out | (1) |
| out | (2) |
| (stops) | |

Note that the counts in parentheses are cumulative, not incremental. This makes simpler such calculations as these for S2:

Reduced count = 10 - (2+1) = 7

Hinge (original) ranks: 2+1 = 3
2+2 = 4

Such calculations naturally involve cumulative counts.

I) SOURCE

A. R. Cushny and J. R. Peebles (1904). Journal of Physiology.

Other users include. Student (1908). The probable error of a mean. Biometrika 6: 1-25 (used at p. 20).

R. A. Fisher (1925) 1946). Statistical Methods for Research Workers (10th ed.) Oliver and Boyd, 354 pp. (used at p. 121).

exhibit 5 of chapter 6

Condensed version of the calculation of exhibit 4

A) STEM-and-LEAF -- unit = 0.1

```
4|6
3|
2|4
1|023348
0|08
```

B) CONDENSED CALCULATION

```
10|U
3H|1.0
3H|1.8

dh|0.8 up 0.4
S/T|0.2 /-0.2
S/T|2.6 /3.0
out|0.0 /
out|/4.6
```

```
9|T1
3H|1.0
4H|1.4

dh|0.4 up 0.2
T|0.4
T|2.0
det|0.0  (0.8)
det|2.4  (1.8)
```

```
7|T2
3H|1.0
4H|1.4

dh|0.4 up 0.2
T|0.4
T|2.0
det|no more
det|no more
```

```
7|T = T2
H|1.0
M|1.3
H|1.4
tri|1.25
```

```
8|S1
3.5H|1.1
3.5H|1.6

dh|0.5
S|0.6
S|2.1
out|no more
out|2.4
```

```
7|S2
3H|1.0
4H|1.4

dh|0.4
S|0.6
S|1.6
out|no more
out|no more
```

```
7|S = S2
H|1.0
M|1.3
H|1.4
tri|1.25
```

Notice the simple marking used for the various steps:

"U" for the unskipped analysis.

"T1", "T2", -- for the various steps of the T-analysis, with "T" used for the final T-step.

"S1", "S2", -- for the various steps of the S-analysis, with "S" used for the final S-step.

Exhibit 6 shows the schematic plots. Continuing the analysis clearly can have a substantial effect.

Even when the first step of T-analysis does not go as far as the first step of the S-analysis, it may happen, as this example illustrates, that the final steps reach the same result. This is, of course, not always so.

Having looked at the revealing schematic plot, we can usefully look at the trimeans, which are:

| | | |
|---|---|---|
| 1.35 | for | "U" |
| 1.25 | for | "T1" |
| 1.25 | for | "T2 = T" |
| 1.32 | for | "S1" |
| 1.25 | for | "S2 = S" |

No basic changes, but no indication of going the wrong way.

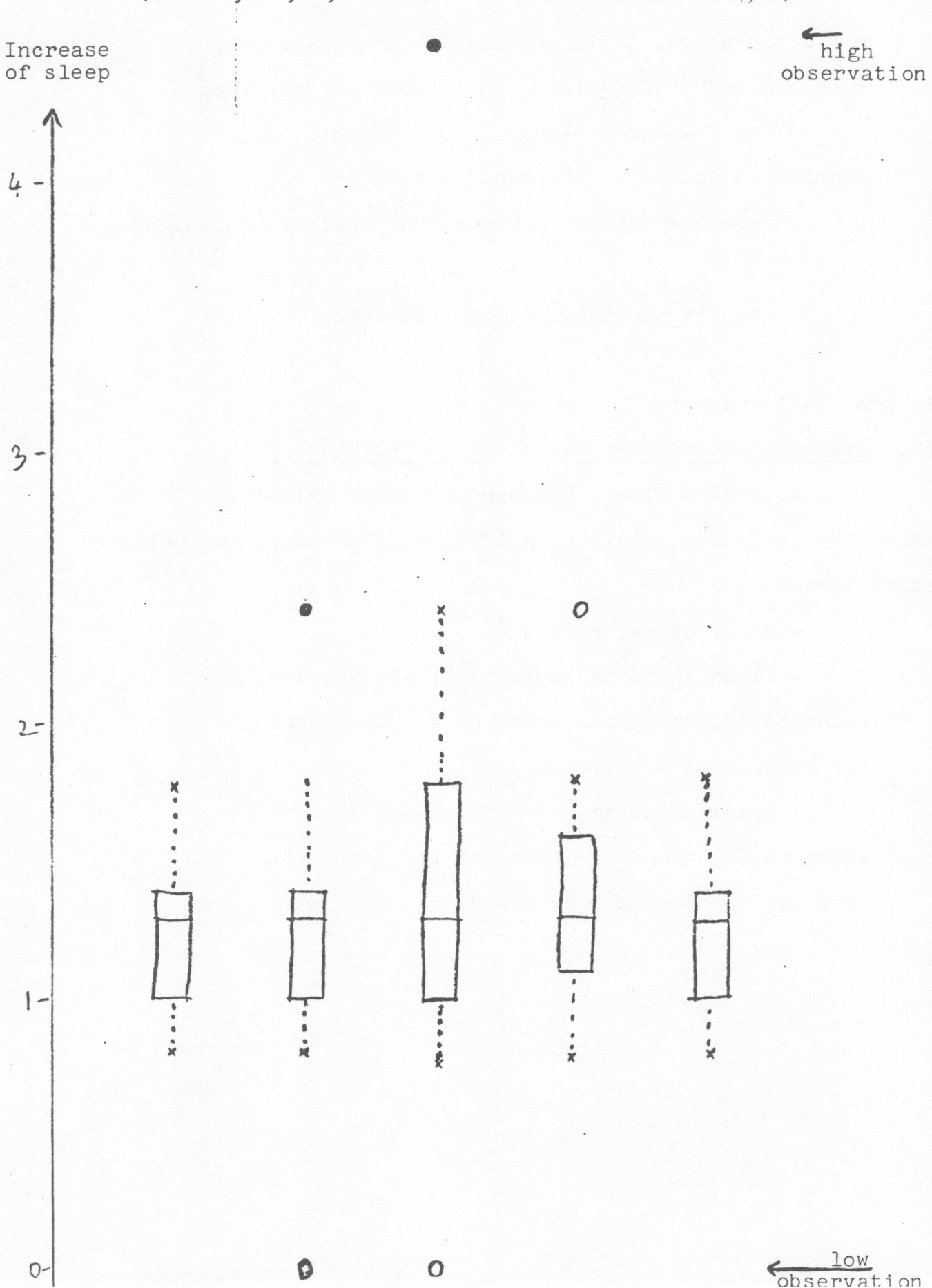
In setting up a multiple skipping process we need to notice that it can come to an end in various ways. Among these are:

- the S or T values repeat, as in panels B and C of exhibit 4.

- the S -- or T -- values may shrink not repeat, but no new outside -- or detached -- observations

Five analyses of the Cushing and Peebles Data

(T = T2, T1, U, S1 and S2 = S from Left to Right)



may be revealed, as illustrated on the low side by panels E and F of exhibit 4. (We may as well stop.)

- the S -- or T -- values may expand, as in the example to follow. (start of section 6F)

In whatever way it happens, the analysis in which

no new observations are set aside

is the final analysis.

6E. Standard patterns of skipping; u-, t-, and s-

We need a simple standard scale of skipping. Three levels -- including no skipping, seem quite enough. We shall adopt these:

- no skipping, marked "u".
- light skipping marked "t", and consisting of going to TF, the final stage of the T-process. (Not used for batches of 4 or fewer.)
- heavy skipping, marked "s" and consisting of going to Sf, the final stage of the S process. (Not used for batches of 2 or fewer.)

For the Cushing and Peebles data, then, we have

u-trimean = 1.35

t-trimean = 1.25

s-trimean = 1.25

Since we do expect fairly frequent stray values in data of this kind, the natural advance decision would be to use either light or heavy skipping with this data. Since either gives a trimean of 1.25, there is no further problem.

Similarly, we have

u-midspread = 0.8

t-midspread = 0.4

s-midspread = 0.4

again there is no difficulty of choice. There is, however, a need for caution. Setting-aside almost inevitably reduces spread. In thinking about the value of t-midspread or the s-midspread it is usually urgent to recall that there has been skipping. (In thinking about the value of the t-trimean or s-trimean, this is not the case.)

We apply the prefixes "u-", "t-" and "s-" freely to any named quantity. Without exception, they mean merely "calculated for the corresponding batch, either as given (for "u-") or after the corresponding set-aside (for t- or s-). Thus skipping midmeans and skipping ranges, for example, are defined as a routine matter (as are skipping midsquares, to be defined when we reach chapter 7y).

Either skipping trimeans or skipping midmeans give us a highly resistant summary of where our values are "centered". Skipping midspreads (or the skipping mid-squares) give us equally resistant measures of spread.

We also need to be aware how easy it is to present all three standard analyses in one combined schematic plot. Exhibit **7** shows the way.

6F. Further examples

(* hydroelectric power again *)

Exhibit 1 of chapter 3 gave log ultimate power for 34 hydroelectric power plants of the Bureau of Reclamation. The corresponding condensed summaries are:

| | | | | | |
|---|---|---|---|---|---|
| 34 | U | 31 | S1 | 33 | T1 |
| 9H | 1.52 | 8H | 1.51 | 9H | 1.52 |
| 17.5M | 1.86 | 16M | 1.78 | 17M | 1.85 |
| 9H | 2.21 | 11H | 2.13 | 10H | 2.18 |
| tri | 1.88 | tri | 1.80 | tri | 1.85 |
| dh | 0.69 up 0.34 | dh | 0.62 | dh | 0.66 up 0.33 |
| (1.18) S/T | 0.83 / 0.48 | S | 0.89 | T | 0.53 |
| (2.63) S/T | 2.90 / 3.25 | S | 2.75 | T | 3.17 |
| out | none / none | out | no more | out | no more |
| out | 2.93, 3.12 / 3.30 | out | no more | out | no more |

The initial and skipping analyses are shown in exhibit **8**. By returning to the source from which the data came, we can meet part of our obligation by identifying the three power plants whose ultimate powers were found to be outside. As one might expect, all 3 are familiar names. We would, however, have to know much more about hydroelectric power possibilities than we do to go further, asking whether -- and in what ways --

exhibit 7 of chapter 6

Combined schematic plot for u-, t- and s-analyses

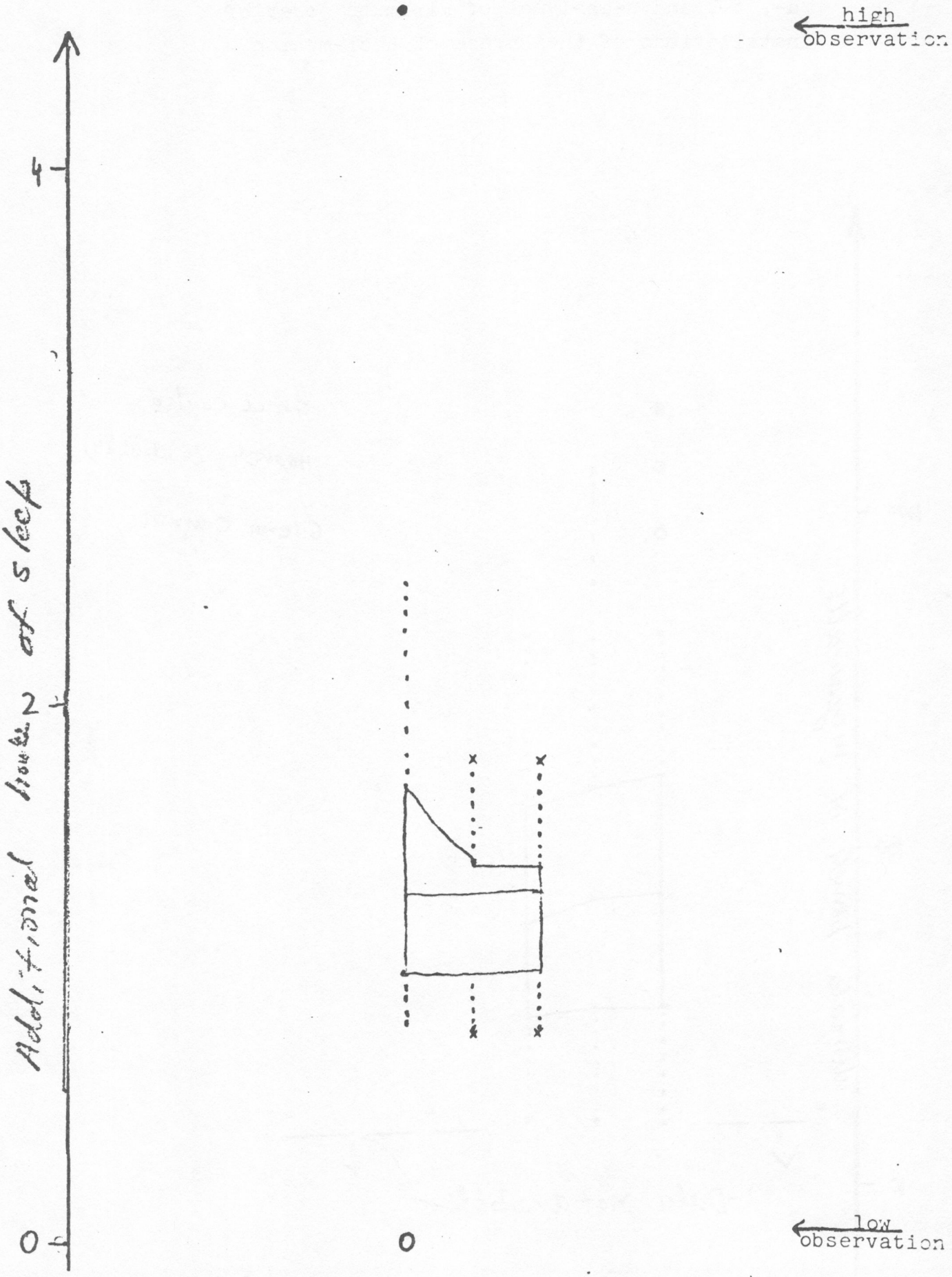
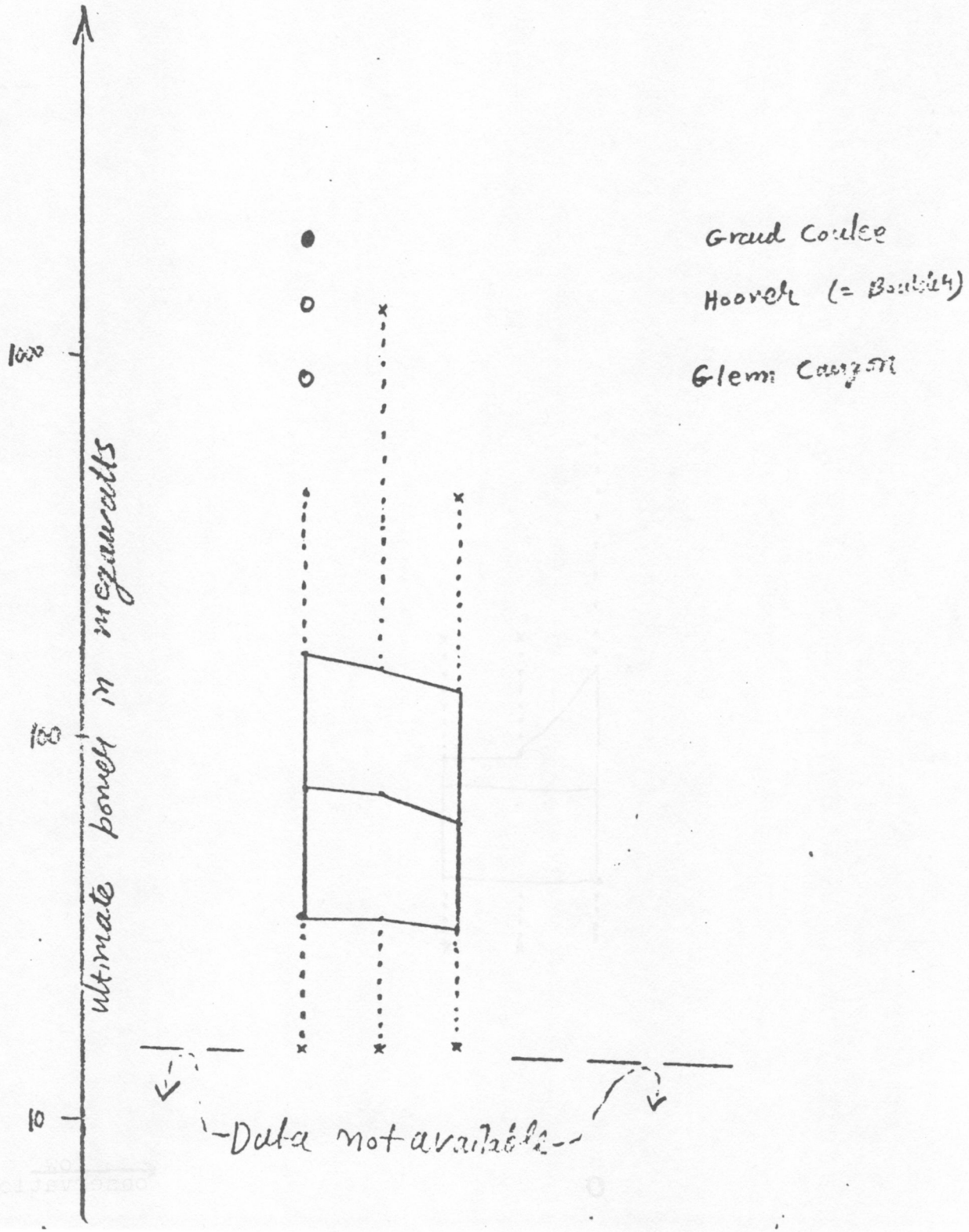


exhibit 8 of chapter 6

u-, t-, and s-analyses of ultimate power of installations of the Bureau of Reclamation



these three locations are to be reasonably distinguished from other locations of the Bureau of Reclamation.

Exhibit 4 of chapter 3 gives log ultimate powers for 62 hydroelectric plants of the Corps of Engineers. For this example, the s-analysis can be repeated more than twice. Exhibit **9** gives the successive condensed notations. In this instance, there were 8 successive s-analyses, rather a large number.

On the other hand, the T-analysis found no detached values, so that all t-values are the same as u-values.

Exhibit **10** shows the combined schematic plot, identifying the installations with the largest and 2 smallest ultimate powers. With a total of 15 high and 4 low values set-aside by the s-analysis, it is even more important to identify and study at least the extreme installations.

When we trace down the 6 largest, we find that all of them occur in Oregon or Washington on the Columbia River or on its tributary, the Snake. Of the 62 installations contributing to exhibit 4 of chapter 3, 8 are on the Columbia or the Snake. Exhibit **11** lists them, and gives a stem-and-leaf for the remaining 54.

This time the s-analysis takes only 4 steps, but it stops in exactly the same place -- hinges and median -- as the previous analysis stopped after 8 steps. If we wish to set aside apparently possibly strays it is clearly far from enough to separate out the Columbia basin. The

The successive skipping analyses for log
ultimate power of Corps of Engineers installations
(data in exhibit 4 of chapter 3)

| | |
|---|---|
| 62 | U |
| 16H | 1.83 |
| 1.5M | 2.04 |
| 16H | 2.45 |
| tri | 2.09 |
| dh | 0.62 up 0.31 |
| S/T | 1.21 /0.90 |
| S/T | 3.07 /3.38 |
| out | 1.15 / none |
| out | 3.24, 3.24, 3.15 / 3.44 |
| adj | 1.26, 2.91 |

| | |
|---|---|
| 57 | S1 |
| 16H | 1.83 |
| 30M | 2.00 |
| 19H | 2.31 |
| tri | 2.04 |
| dh | 0.48 |
| S | 1.35 |
| S | 2.79 |
| out | 1.26 |
| out | 2.91, 2.91 |
| adj | 1.42, 2.78 |

| | |
|---|---|
| 54 | S2 |
| 16H | 1.83 |
| 29.5M | 2.00 |
| 20H | 2.30 |
| tri | 2.03 |
| dh | 0.47 |
| S | 1.36 |
| S | 2.77 |
| out | none |
| out | 2.78 |

| | |
|---|---|
| 53 | S3 |
| 16H | 1.83 |
| 29M | 2.00 |
| 21H | 2.24 |
| tri | 2.02 |
| dh | 0.41 |
| S | 1. 2 |
| S | 2.65 |
| out | none |
| out | 2.77,2.73,2.71, 2.67 |

| | |
|---|---|
| 47 | S4 |
| 15H | 1.79 |
| 27M | 2.00 |
| 23H | 2.13 |
| tri | 1.98 |
| dh | 0.34 |
| S | 1.45 |
| S | 2.47 |
| out | 1.42 |
| out | 2.60, 2.53, 2.52,2.51 |

| | |
|---|---|
| 44 | S5 |
| 145H | 1.7 |
| 25.5M | 2.00 |
| 26.5H | 2.1 |
| tri | 1.9 |
| dh | 0.33 |
| S | 1.45 |
| S | 2.44 |
| out | none |
| out | 2.45 |

| | |
|---|---|
| 43 | S6=S |
| 14H | 1.78 |
| 25M | 2.00 |
| 27H | 2.11 |
| tri | 1.97 |
| dh | 0.33 |
| S | 1.45 |
| S | 2.44 |
| out | none |
| out | none |
| | stops |
| adj | 1.45, 2.43 |

Note: The value of listing individual outside observations decreases greatly as soon as we have many steps. Had we used the other style, and entered counts after out , we would have found the calculation simpler. The entries would have been, in order:

| | | | | | | |
|---|---|---|---|---|---|---|
| (1,0) | (2) | (2) | (2) | (3) | (3) | (3) |
| (3,1) | (6) | (7) | (11) | (15) | (16) | (16) |

exhibit 10 of chapter 6

u-, t-, and s-analyses of ultimate power of installations of the Corps of Engineers

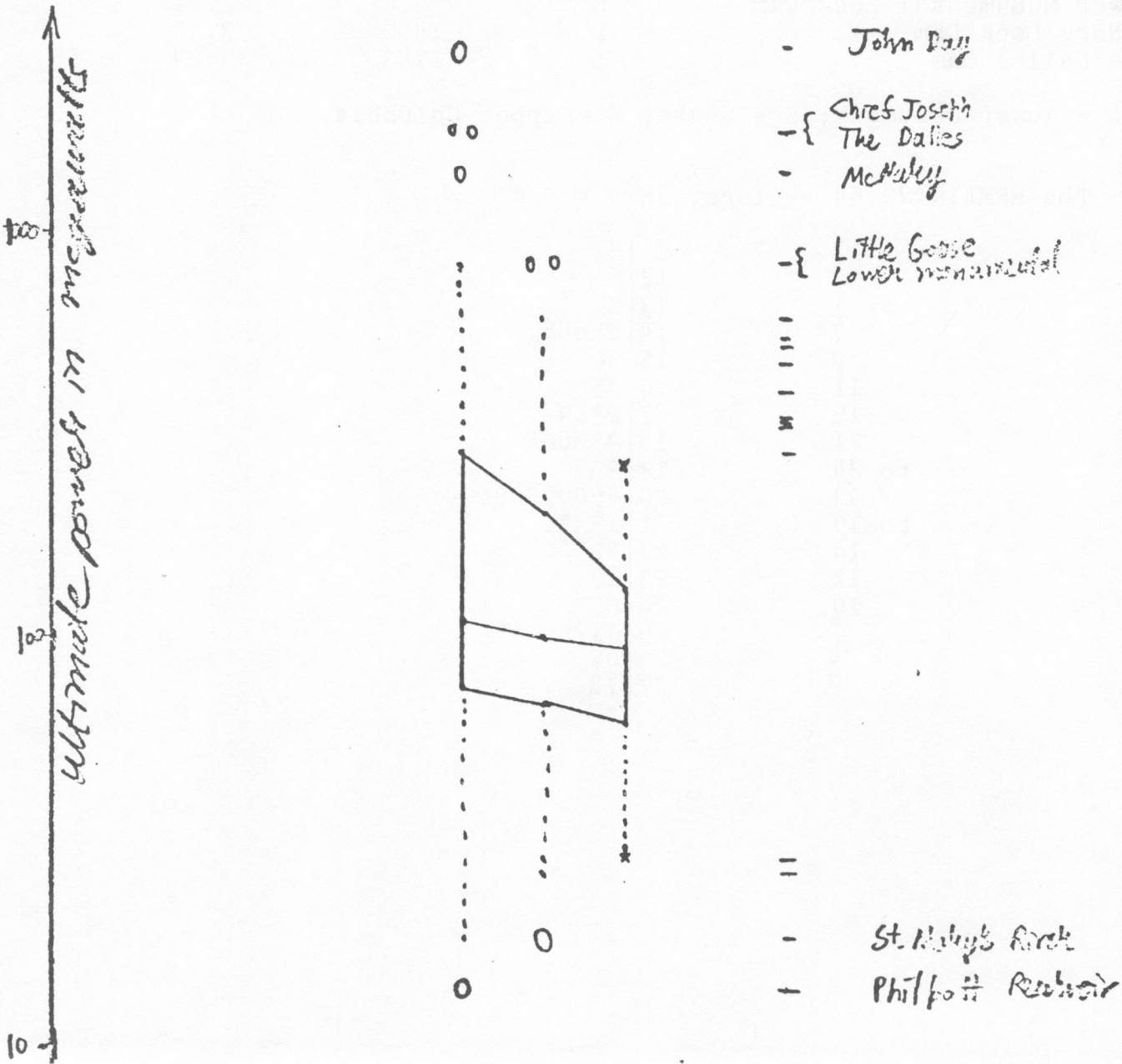


exhibit 11 of chapter 6

Corps of Engineers hydroelectric installations sorted out

(see exhibit 4 of chapter 3)

**A)** On the COLUMBIA or the SNAKE

| Name | * | Ultimate Power (megawatts) Raw | Log |
|---|---|---|---|
| Bonneville Dam | L | 518 | 2.71 |
| Chief Joseph Dam | U | 1728 | 3.24 |
| Ice Harbor Lock Dam | S | 540 | 2.73 |
| John Day Lock Dam | C | 2700 | 3.43 |
| Little Goose | S | 810 | 2.91 |
| Lower Monumental Lock Dam | S | 810 | 2.91 |
| McNary Lock Dam | L | 1400 | 3.15 |
| The Dalles Dam | L | 1743 | 3.24 |

* L = lower Columbia; S = Snake; U = upper Columbia.

**B)** The REMAINING 54 -- logs, unit = 0.01

| | | |
|---|---|---|
| 1 | 11 | 5 |
| 2 | 12 | 6 |
| . | 13 | |
| 7 | 14 | 25888 |
| 9 | 15 | 36 |
| 11 | 16 | 35 |
| 15 | 17 | 2389 |
| 21 | 18 | 335888 |
| to 24 | 19 | 358 |
| ✓ 11 | 20 | 45009700004 |
| to 19 | 21 | 13351 |
| 14 | 22 | 42 |
| 12 | 23 | 10 |
| 10 | 24 | 053 |
| 7 | 25 | 312 |
| 4 | 26 | 07 |
| 2 | 27 | 78 |

exhibit 11 of chapter 6 (Cont'd)

C) ANALYSES of the 54

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 54 | U | 49 | S1 | 44 | S2 | 42 | S3 = S |
| 14H | 1.78 | 15H | 1.79 | 14.5H | 1.785 | 15H | 1.79 |
| 27.5M | 2.00 | 27M | 2.00 | 25.5M | 2.00 | 25.5M | 2.00 |
| 14H | 2.22 | 16H | 2.13 | 18.5H | 2.11 | 19H | 2.11 |
| tri | 2.00 | tri | 1.98 | tri | 1.98 | tri | 1.98 |
| dh | 0.44 up 0.22 | dh | 0.34 | dh | 0.325 | dh | 0.32 |
| ST | 1.34 /1.12 | S | 1.45 | S | 1.46 | S | 1.47 |
| ST | 2.66 /2.88 | S | 2.47 | S | 2.435 | S | 2.43 |
| out | (2,0) | out | (3) | out | (4) | out | no more |
| out | (3,0) | out | (7) | out | (8) | out | no more |

pattern of ultimate powers for Corps of Engineers hydro-electric insta-lations encourages us to set aside 12 more installations, 8 as high and 4 as low.

The t-values are again the same as the u-values.

So long as we set aside the Columbia basin, however, the trimean lies between 2.00 and 1.98. This can be compared with the figures for the Bureau of Reclamation installations, provided Grand Coulee, on the upper Columbia, is removed.

The condensed analysis for the 33 powers left runs:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | U | dh | 0.66 up 0.33 | / | 31 | S1 | dh | 0.62 |
| 9H | 1.52 | S/T | 0.86 / 0.53 | 9 | 8H | 1.51 | S | 0.89 |
| 17M | 1.85 | S/T | 2.84 / 3.27 | 9 | 16M | 1.78 | S | 2.75 |
| 9H | 2.18 | out | none / none | 9 | 10H | 2.13 | out | none |
| tri | 1.85 | out | 3.12, 2.95 / none | 9 | tri | 1.80 | out | none |

The difference in trimeans of logs is, thus:

2.00 - 1.85 = 0.15 u-trimean

1.98 - 1.85 = 0.13 t-trimean

1.98 - 1.80 = 0.18 s-trimean

Even if we set aside the Columbia basin, the Corps of Engineers installations seem to run about 30% to 50% larger than those of the Bureau of Reclamation (log 1.3 = .12, log 1.5 = .18).

(* atomic weights *)

Exhibit 15 of chapter 3 gave two batches of atomic weight determinations, one for hydrogen and one for bromine. Exhibit 12 does a skipping analysis. The three versions

Skipping analysis of atomic weights

A) HYDROGEN -- excess of logs over .0033, unit .000001

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 9 | U | 8 | T | 7 | S | inside S |
| 2 | 7 | 3H | 53 | 3.5H | 58 | 4H | 62 | |
| 3 | | 5M | 70 | 5.5M | 70 | 6M | 70 | |
| 4 | | 3H | 78 | 2.5H | 80 | 2H | 82 | |
| 5 | 3 | tri | 68 | tri | 70 | tri | 71 | 3 |
| 6 | 2 | | | | | | | 2 |
| 7 | 080 | dh | 25 up 12 | dh | 22 up 11 | dh | 20 | 080 |
| 8 | 27 | ST | 28 /16 | T | 25 | S | 42 | 27 |
| | | ST | 103 /115 | T | 113 | S | 102 | |
| | | out | 27 / 14 | out | no more | out | no more | |
| | | out | none / none | out | no more | out | no more | |

(mean = .003360)

u-trimean = .003368

t-trimean = .003370

s-trimean = .003371

B) BROMINE -- excess of logs over .19026, unit .0000001

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| -0 | 4 | 10 | U | 9 | T | 7 | S | inside S |
| 0 | | 3H | 28 | 3H | 28 | 3H | 28 | |
| 1 | 2 | 5.5M | 40 | 5M | 34 | 5M | 34 | 2 |
| 2 | 88 | 3H | 56 | 4H | 56 | 4H | 56 | 88 |
| 3 | 4 | tri | 41 | tri | 38 | tri | 38 | 4 |
| 4 | 5 | | | | | | | 5 |
| 5 | 66 | dh | 28 up 14 | dh | same | dh | same | 66 |
| 6 | | S/T | 00 /-14 | T | | S | | |
| 7 | | S/T | 84 / 98 | T | | S | | |
| 8 | 8 | out | -04 / none | out | | out | | |
| 9 | | out | 88 / 148 | out | | out | | |
| 10 | | | | | | | | |
| 11 | | | | | | | | |
| 12 | | | | | | | | |
| 13 | | | | | | | | |
| 14 | 8 | | | | | | | |

(mean = .1902649)

u-trimean = .1902641

t-trimean = .1902638

s-trimean = .190263

correspond well to our intuitive ideas as to what are reasonable analyses of these numbers. The small changes in trimean reflect credit on the resistance of the initial trimeans. However, the values set aside appear rather clearly discrepant.

That some determinations appear clearly discrepant is probably an indication they are either worse or better than the others. The obligation on a chemist at this stage of the analysis would be to go back and read the original papers. Then he might be able to apply his judgment wisely.

## 6G. Generations in data analysis

Today it is fashionable to distinguish "generations" in inanimate objects -- as when we speak of third-generation computers, for example. Not only fashionable but sometimes helpful.

The use of middle means is, in this sense, third-generation data analysis. First-generation methods acted as if all values were well-behaved. This led to using the mean for a central value -- with good effects when everything was indeed well-behaved. Rather too often, however, usually where violently straying values occurred, the harvest was confusion and error.

In over-reaction to this, second-generation methods assumed every observation to be ill-behaved, and sought as much protection from ill-behavior as possible. One result was using the median for a central value. This gave extreme protection against confusion and error, but cost somewhat more than necessary when all values happened to be well-behaved.

The return swing of the pendulum was shorter. Third-generation methods anticipate a mixture of well-behaved and ill-behaved values. Experience teaches us it is realistic to do just this.

The midmean is simple, and deals quite well with very small numbers of widely straying values. If we had to criticize it, we would have to criticize it from two, apparently opposed directions:

- it is unnecessarily sensitive to very widely straying values, since their existence and number -- though not their size -- affect which values are to be hinges.

- it is unnecessarily restrictive, especially for small batches, since it is often clear that the hinge values can perfectly well be included with the middle values before we form the mean.

The trimean is equally subject to the first of these criticisms -- and not resistant enough in batches of 3 or 4.

The skipping forms of the trimean and midmean meet these criticisms quite thoroughly.

In a real and useful sense skipping trimeans and skipping midmeans are much more nearly the full flower of third-generation technique than their unskipped relatives.

*****

6H. How far have we come?

In this chapter we have learned to do skipping analyses. To do this we use a set-aside procedure, sometimes once, but usually as many times as possible.

Once we have finished setting aside, we plan to summarize the remaining values along the same lines as before:

- with a trimean or midmean,

and, if a measure of spread is desired:

- with a midspread (or after section 7yA, a midsquare).

The only novel feature here is including the hinge values (if not interpolated) with the middle values in forming the midmean.

We also plan to report outside and detached observations, identifying them if we can, and apply judgment to them.

The set-aside procedure itself is simple. Just repeated application of a simple process:

- find hinges, and either side values, or three-halves values, identifying observations outside these as either outside observations or detached observations.
- set aside these outside or detached observations.

Two standard extents of repeated set-aside are:

- s-skipping, in which we set aside outside points as often as new ones appear.
- t-skipping, in which we set aside detached observations as often as new ones appear.

For batches of 3 or 4 we have learned special ways to set side values. (Side points 10 times or twice, respectively, as far outside each observation of depth 2 as that observation is from the opposite extreme.)

Finally, we learned that we should expect to often look at two alternate analyses -- if not more.

*****

6K. How do the skipping midmeans and trimeans respond to changing one value in a batch? (background)

In section 4K we looked to see what happened to various central values when we shifted one value of Lord Rayleigh's seven weights of nitrogen from air. Let us look, in the same way, at what happens when we use the skipping trimean or skipping midmean. Exhibit 13 shows the behavior of the skipping midmean when the first of the seven values is shifted. The behaviors of mean, median, and midmean are shown as background. The relapse of the skipping midmean values to 2.310157, when the first value is made either quite large or quite small is obviously almost pure profit. Exhibit 14 shows the same thing for the skipping trimean.

6L. Skipmeans (optional technique)

Once we use skipping, we have an opportunity to re-examine one of our earlier choices. Back in chapter 4 -- in section 4F where we discussed the choice between trimean and midmean -- we emphasized the greater resistance of the midmean, especially in small batches. This greater resistance came because the hinges were not included among the observations whose mean is the midmean.

Once we use skipping, we are protected against very stray observations because they will be set aside. Hence we do not need as much resistance in the central value to be calculated from the remaining observations.

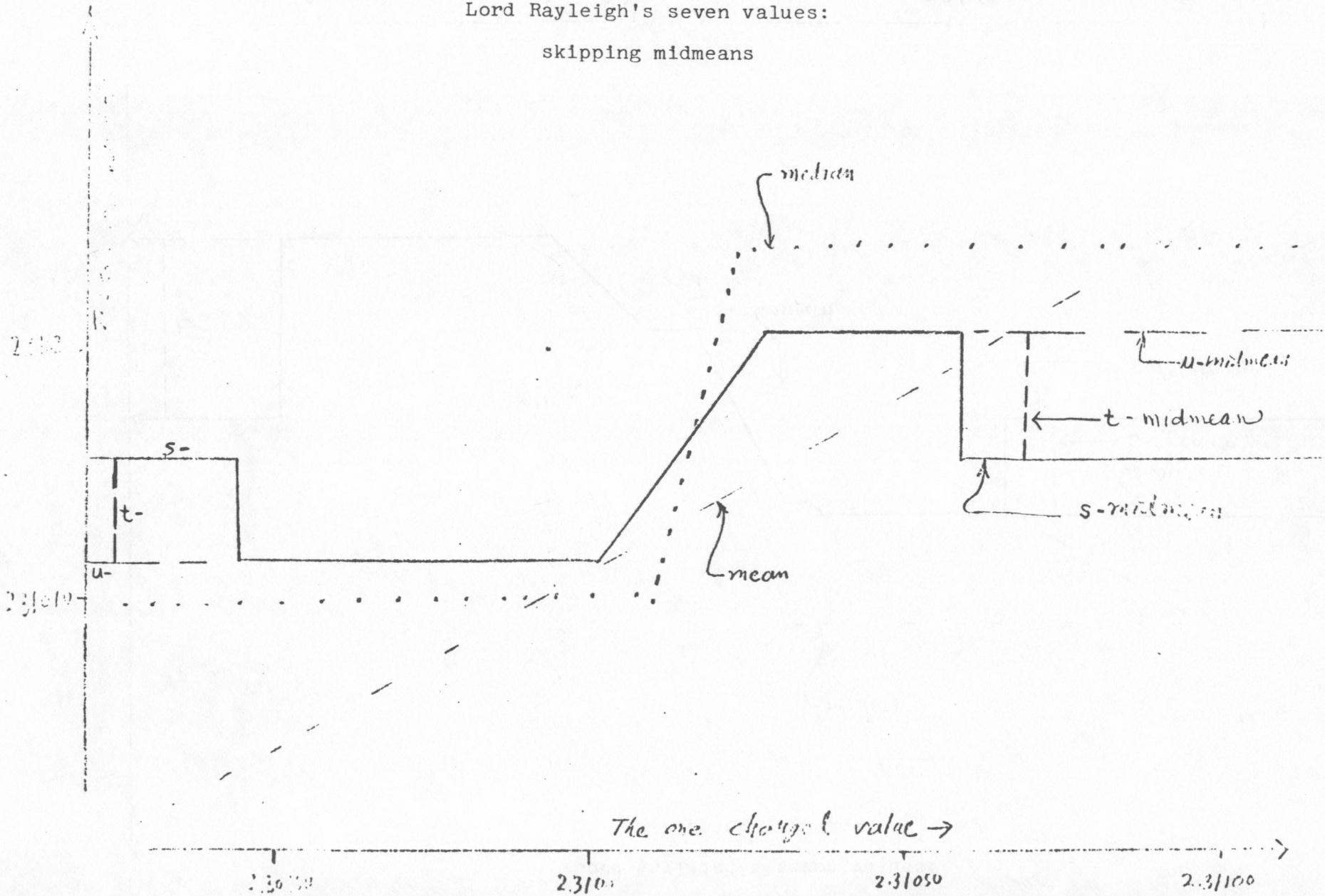
It is not unreasonable to consider using the

skipmean

The result of moving the first of

Lord Rayleigh's seven values:

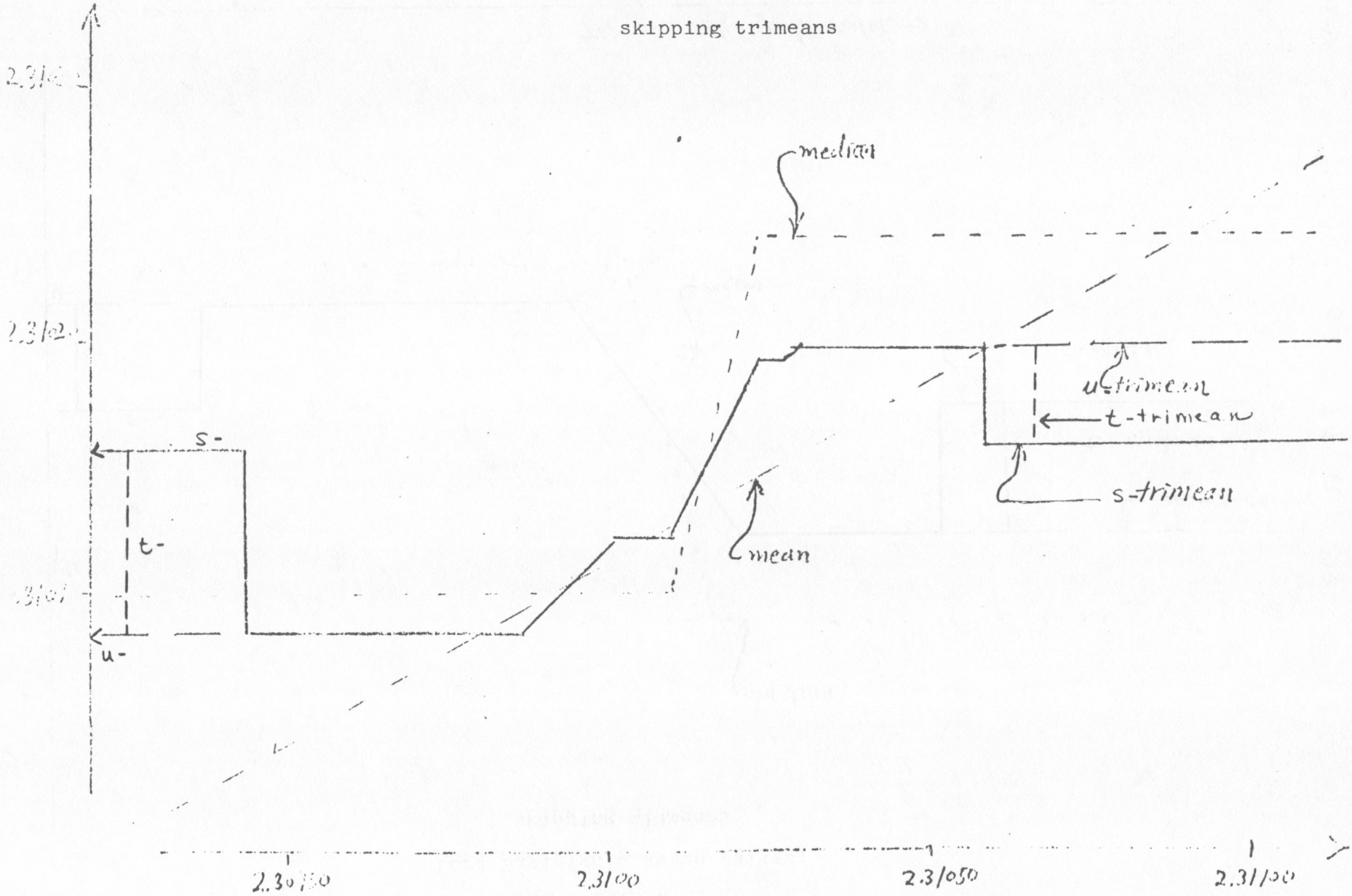
skipping midmeans



The result of moving the first of

Lord Rayleigh's seven values:

skipping trimeans



defined -- in terms of the hinges of the batch remaining after s-set aside -- as the mean of the hinges and all the observations between the hinges. For the 7 Rayleigh values with one changed, the behavior of the skipmean is as shown in exhibit 15. In this specific case, the skipmean seems to behave slightly better than the s-midmean and appreciably better than the s-trimean.

There seem to be many cases where using the skipmean produces slightly better results. The gain is usually small, however, and it is not yet clear whether it is worth the complication of introducing an additional central value.

Exhibit 16 gives the detailed numbers for moving each of the seven values in turn. They offer no reason to change our belief that the skipping versions have many desirable properties. Recall that in section 3K we felt the overall change in a central value as an extreme stray went from very low to very high was one thing we wanted to make smaller. Skipping does not make the largest swing any smaller, since the largest swings arise before set-aside can have any effect.

6M. <u>s-means</u>

We ought also to note that it is sometimes useful to take the mean of the values remaining after skipping. The result is clearly the

s-mean .

The result of moving the first of

Lord Rayleigh's seven values

skipmeans

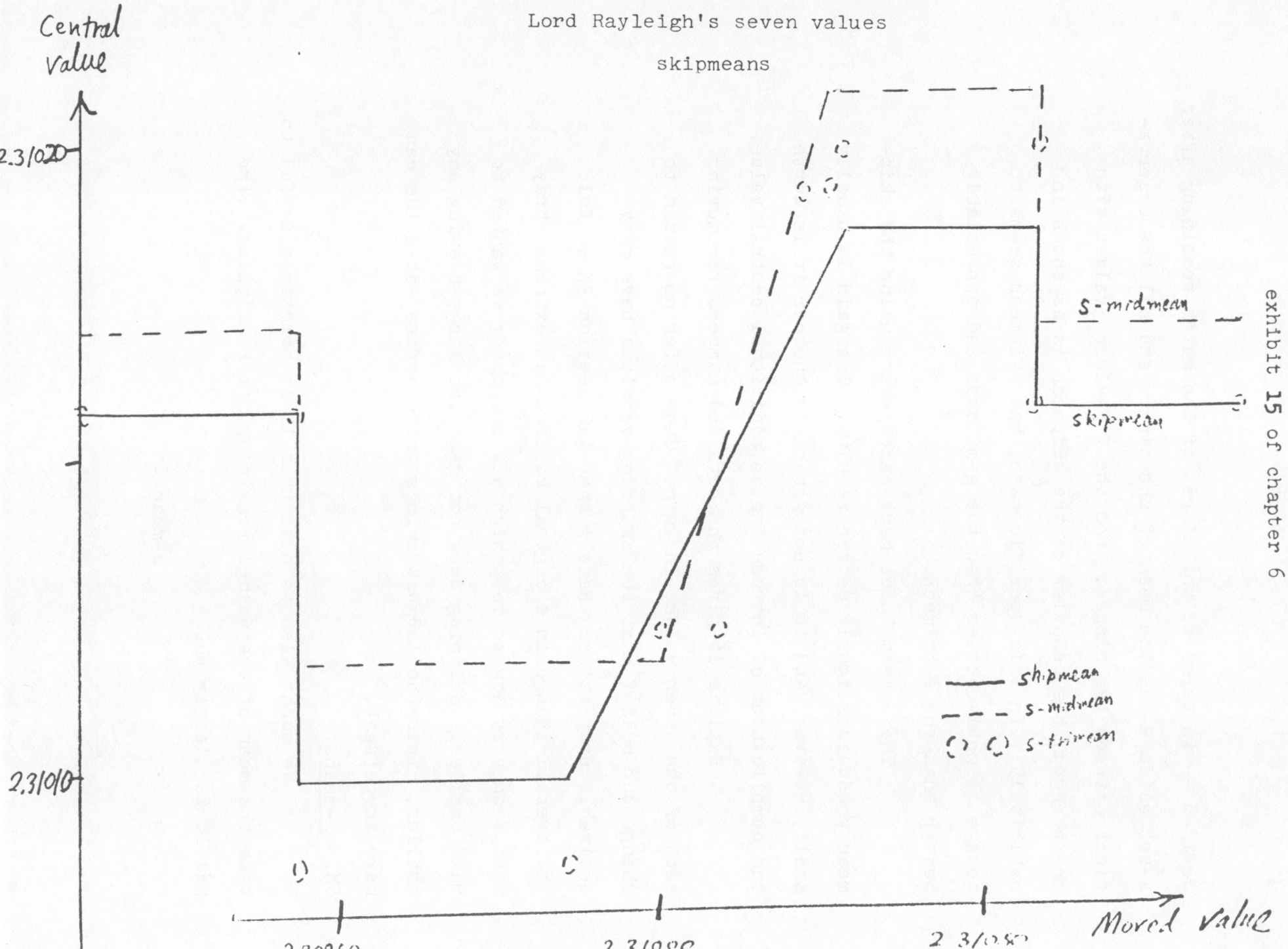



exhibit 15 of chapter 6

exhibit 16 of chapter 6

The effects of moving each of Lord Rayleigh's seven values on skipping midmean and skipping trimean

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Value moved: | -14 | 1 | 10 | 17 | 24 | 28 | 30 |
| Values* that cause set aside | -56<br>50 | -56<br>50 | -56<br>59 | -56<br>59 | -56<br>59 | -52<br>59 | -52<br>55 |
| midmean, if set aside occurs | 19.8 | 19.8 | 17.5 | 15.7 | 14 | 13 | 13 |
| trimean, if set aside occurs | 19.8 | 19.8 | 17.5 | 15.7 | 14 | 13 | 13 |

* Innermost values for the moved values that produce set aside.

Note: s-midmean and s-trimean when one value is set aside will usually be unequal, even though they are always equal when, as here, exactly one value out of exactly seven is set aside.

*****

6P. Problems

(1) Make a combined schematic plot comparing, according to u-, t-, and s-analyses, the ultimate powers of hydroelectric installations not on the Columbia or Snake Rivers as between the Bureau of Reclamation and the Corps of Engineers. (See exhibits 7 to 10 and their sources.) Compare with exhibit 11 of chapter 5. Comment in detail.

* * *

(2) Apply the set-aside procedure to the pooled residuals of panel D of exhibit 3 of chapter 4. Find the u-, t-, and s-trimeans and midspreads.

(3) Do the same for the pooled residuals on the right side of exhibit 13 of chapter 4. (The left side repeats the data analyzed in the previous problem.) Compare these results with those for the left side.

(4) In 1946, Fornachon gave the acidity values for wine in exhibit 17. Calculate the u-, t-, and s-trimeans and midspreads.

(5) In 1894, Loomis reviewed the literature of the freezing point of 3.3% sugar solution (in water), giving the values in exhibit 18. Calculate the u=, t-, and s-trimeans and midspreads.

(6) Do the same for the speeds of sound in exhibit 19.

(7) Do the same for the viscosities of exhibit 12 of chapter 1.

(8) Do the same for the radial velocities of exhibit 13 of chapter 1.

(9) Do the same for the association constants of exhibit 14 of chapter 1.

(10) Do the same for at least 3 of the batches of exhibit 15 of chapter 1.

(11) Do the same for the 4 batches of gold contents in exhibit 8 in chapter 2.

(12) Do the same for the 2 batches of snowfalls in problems (1) and (2) of section 4P.

(13) Do the same for the carotene percents of problem (6) of section 4P.

(14) Do the same for the drape measurements of exhibit 18 of chapter 4.

(15) Do the same for the Kjeldahl ultramicrodeterminations of exhibit 22 of chapter 4.

(16) Do the same for at least 3 of the batches of critical temperatures in exhibit 21 in chapter 5.

exhibit 17 of chapter 6

Some pH's for Wines

A) The pH VALUES -- glass electrode

| Sweet red wine | | Dry white wine | |
|---|---|---|---|
| 3.4 | 5 | 2.9 | 5 |
| 3.5 | | 3.0 | |
| 3.6 | 6944 | 3.1 | |
| 3.7 | 4213 | 3.2 | 2 |
| 3.8 | 4 | 3.3 | 2 |
| | | 3.4 | 0 |
| | | 3.5 | 9 |
| | | 3.6 | |
| | | 3.7 | 3 |
| | | 3.8 | 0 |

B) NOTE

Lomer pH means more acid. plt = 7.0 corresponds to neutral.

C) The SOURCE

J. C. M. Fornachon 1946. The pH of wines. Examination of glass and quinhydrone electrode values. Industrial and Engineering Chemistry, Analytical Edition 18: 790-793. Table 1 on page 791.

exhibit 18 of chapter 6

Freezing points of sugar solutions before 1894

(3.3% sugar in water)

A) The DATA

| Date | Experimenter | Freezing points (°C) |
|---|---|---|
| 1886 | Raoult I | 0.24 |
| 1888 | Arrhenius I | 0.210 |
| 1891 | Traube | 0.235 |
| | Eijkmann | 0.216 |
| | Arrhenius II | 0.204 |
| | Tammann | 0.206 |
| | Pickering | 0.201 |
| 1892 | Raoult II | 0.205 |

B) The SOURCE

E. H. Loomis 1894. Ueber ein exacteres Verfahren bei der Bestimmung von Gefrierpunktserniedriguuyen. Annalen der Physik und Chemie 51: 500-523. Page 517.

exhibit 19 of chapter 6

Various measurements of speed of sound

A) The SPEEDS in AIR -- in meters per second at 0°C.

| open form |
|---|
| 332.77 |
| 331.57 |
| 333 |
| 330.66 |
| 330.71 |
| 332.06 |
| 332.5 |
| 333.898 |
| 330.7 |
| 331.29 |
| 331.676 |
| 331.36 |
| 332.1 |
| 331.5 |
| 331.92 |
| 333.4 |
| 330.8 |
| 331.57 |
| 331.52 |

| mixed stem-and-leaf | |
|---|---|
| 330. | 66,71,7,8, |
| 331. | 29,36, |
| 331. | 57,676,5,92,57,52, |
| 332. | 06,5,1, |
| 332. | 77, |
| 333. | -, 4, |
| 333. | 898 |

B) The SOURCE

Landolt-Bornstein 1923.

5th Air and hydrogen table 338 on page 1631.

B) The SPEED in HYDROGEN -- meters per second at 0°C

1270, 1286, 1259, 1261, 1260.6

C) The SPEED in CARBON DIOXIDE -- meters per second at 0°C